Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 48 (2006) 1458–1485
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Feature extraction from higher-lag autocorrelation coefficients for robust speech recognition

Benjamin J. Shannon, Kuldip K. Paliwal *

School of Microelectronic Engineering, Griffith University, Nathan Campus, Brisbane, QLD 4111, Australia

Received 4 July 2005; received in revised form 1 August 2006; accepted 1 August 2006

Abstract

In this paper, a feature extraction method that is robust to additive background noise is proposed for automatic speech recognition. Since the background noise corrupts the autocorrelation coefficients of the speech signal mostly at the lower-time lags, while the higher-lag autocorrelation coefficients are least affected, this method discards the lower-lag autocorrelation coefficients and uses only the higher-lag autocorrelation coefficients for spectral estimation. The magnitude spectrum of the windowed higher-lag autocorrelation sequence is used here as an estimate of the power spectrum of the speech signal. This power spectral estimate is processed further (like the well-known Mel frequency cepstral coefficient (MFCC) procedure) by the Mel filter bank, log operation and the discrete cosine transform to get the cepstral coefficients. These cepstral coefficients are referred to as the autocorrelation Mel frequency cepstral coefficients (AMFCCs). We evaluate the speech recognition performance of the AMFCC features on the Aurora and the resource management databases and show that they perform as well as the MFCC features for clean speech and their recognition performance is better than the MFCC features for noisy speech. Finally, we show that the AMFCC features perform better than the features derived from the robust linear prediction-based methods for noisy speech.

Keywords: Speech recognition; Feature extraction; Robustness to noise; MFCC

1. Introduction

A speech recogniser is trained in a given acoustic environment and deployed (or, tested) normally in a different environment; thus, there is always a mismatch in the training and test environments. This mismatch causes a drastic degradation in speech recognition performance. One of the major factors responsible for the mismatch between the training and test environments is additive background noise (uncorrelated to speech) (Juang, 1991; Gong, 1995). A number of methods have been proposed in the literature to overcome this environmental mismatch problem. These include robust feature extraction methods (Ghitza, 1986; Mansour and Juang, 1989; Paliwal and Sondhi, 1991; Paliwal and Sagisaka, 1997), speech enhancement methods (Kim et al., 2003; Hermus and Wambacq, 2004), feature compensation methods (Hermansky and Morgan, 1994; Stern et al., 1996), multi-band methods (Bourlard

* Corresponding author. Tel.: +61 7 3875 6536; fax: +61 7 3875 5198.
E-mail address: K.Paliwal@griffith.edu.au (K.K. Paliwal).

0167-6393/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2006.08.003

and Dupont, 1996; Tibrewala and Hermansky, 1997), missing feature methods (Lippmann, 1997; Cooke et al., 1997; Raj et al., 2004), and model compensation (and adaptation) methods (Gales and Woodland, 1996; Bellegarda, 1997; Lee, 1998).

The focus of this paper is on robust feature extraction for speech recognition. We are interested in developing a feature extraction method that can deal with the additive background noise distortion in a robust manner. Here, we are given a frame of the observed (noisy) signal $x(n)$, $n = 0, 1, \ldots, N-1$, for analysis, where N is the frame length (in number samples). This can be expressed as

$$x(n) = s(n) + d(n), \tag{1}$$

where $s(n)$ is the clean speech signal and $d(n)$ is the background noise signal.[1] Our aim is to extract recognition features from the noisy speech signal $x(n)$ in such a manner that they capture the spectral characteristics of the clean speech signal $s(n)$ accurately and are least affected by the noise $d(n)$.

The Mel-frequency cepstral coefficients (MFCCs) are perhaps the most widely used features in the current state-of-the-art speech recognition systems (Rabiner and Juang, 1993; Huang et al., 2001). In MFCC feature extraction (Davis and Mermelstein, 1980), the noisy signal $x(n)$ is processed in terms of the following steps: (1) Perform short-time Fourier analysis of the signal $x(n)$ using a finite-duration window (such as a 32 ms Hamming window) and use the periodogram method (Kay, 1988) to compute the power spectral estimate $\widehat{P}_{xx}(\omega)$ of the signal $x(n)$, (2) apply the Mel filter bank to the power spectrum $\widehat{P}_{xx}(\omega)$ to get the filter-bank energies, and (3) compute discrete cosine transform (DCT) of the log filter-bank energies to get the MFCCs. The MFCC features perform reasonably well for the recognition of clean speech, but their performance is very poor for noisy speech. This happens because the periodogram-based power spectral estimate used in MFCC computation gets severely affected by the additive background noise and this degrades the recognition performance of MFCC features for noisy speech.

In the present paper, we use the autocorrelation domain processing for robust estimation of power spectrum from noisy speech. The autocorrelation function of a signal is related to the signal's power spectrum through the Fourier transform (Kay, 1988) and it has the following two attractive properties: (1) Additivity property: If the two signals are uncorrelated, the autocorrelation function of their sum is equal to the sum of their autocorrelation functions. Thus:

$$r_{xx}(n) = r_{ss}(n) + r_{dd}(n), \tag{2}$$

as $s(n)$ and $d(n)$ are uncorrelated signals. (2) Robustness property: The autocorrelation function of the white random noise signal is zero everywhere except for the zeroth time-lag (Kay, 1979). For broadband noise signals, it is mainly confined to lower-time-lags and is very small or zero for higher-time-lags (Mansour and Juang, 1989). As a result, the additive noise $d(n)$ does not affect the higher-lags of the autocorrelation function. Thus the higher-lag autocorrelation coefficients are relatively robust to additive noise distortion.

Because of these attractive properties, the autocorrelation domain processing has been used in the past for autoregressive (AR) spectral estimation (or, linear prediction (LP) analysis) of the noisy signals. The initial effort in this direction was based on the use of high-order Yule–Walker equations (Gersch, 1970; Chan and Langford, 1982), where the autocorrelation coefficients that are involved in the Yule–Walker equation set exclude the zero-lag coefficient. Other similar methods have been used that either avoid the zero-lag coefficient (Cadzow, 1982; Paliwal, 1986a; Paliwal, 1986b; Paliwal, 1986c; McGinn and Johnson, 1989), or reduce the contribution of the first few coefficients (Mansour and Juang, 1989; Hernando and Nadeu, 1997). All of these techniques are based on all-pole modelling of the causal part of the autocorrelation sequence of signal $x(n)$. Two of these techniques (Mansour and Juang, 1989; Hernando and Nadeu, 1997) have been used to extract the cepstral coefficient features for speech recognition and found to provide some robustness to noise, but their recognition performance for clean speech is worse than the conventional linear prediction cepstral coefficient (LPCC) features (Hernando and Nadeu, 1997).

In the present paper, we propose a robust feature extraction method based on autocorrelation domain processing.[2] Since the broadband noise distortion

[1] We denote the power spectrum of the noisy signal $x(n)$ by $P_{xx}(\omega)$ and its autocorrelation function by $r_{xx}(n)$. Similarly, for the clean signal $s(n)$, the corresponding symbols are $P_{ss}(\omega)$ and $r_{ss}(n)$, respectively. For the noise signal $d(n)$, these symbols are $P_{dd}(\omega)$ and $r_{dd}(n)$, respectively.

[2] We have reported the preliminary results about this method earlier in conferences (Shannon and Paliwal, 2004; Shannon and Paliwal, 2005).